Agribusiness Time Series Forecasting using Perceptually Important Events

Lucas S. Rodrigues
Laboratory of Scientific Computing (LivES)
Federal University of Mato Grosso do Sul
Três Lagoas, MS, Brazil
Email: lucas.rodrigues@alunos.ufms.br

Solange O. Rezende
Institute of Mathematical and Computer Sciences
University of São Paulo
São Carlos, SP, Brazil
Email: solange@icmc.usp.br

Maria F. Moura
Laboratory of Computational Intelligence
Brazilian Agricultural Research Corporation (Embrapa)
Campinas, SP, Brazil
Email: maria-fernanda.moura@embrapa.br

Ricardo M. Marcacini
Laboratory of Scientific Computing (LivES)
Federal University of Mato Grosso do Sul
Três Lagoas, MS, Brazil
Email: ricardo.marcacini@ufms.br

***Resumo*—Modern agribusiness management incorporates instruments for risk management with the objective of mitigating uncertainties to the producer. In this context, the producer (risk averse) transfer the risk of price oscillation to companies or individuals that operate in the futures market and who expect to receive a payment (risk premium) for assuming such risk. Defining the adequate strategies for risk management depends on the knowledge about the problem to determine prices ranges in the future. Recent studies demonstrate that time series forecasting can be significantly improved by considering additional information about the problem. In particular, besides the historical time series, textual knowledge extracted from the news portals, social networking and other public data sources available in the web may also be used. This paper presents an approach for agribusiness time series forecasting that allows incorporating external knowledge in the form of events extracted from news about agribusiness, without the need to previously label textual information. In this case, periods of significant uptrends and downtrends of time series are automatically identified — known in the literature as perceptually important points (PIP). We extend the concept of PIP to news events, where similar events published with a certain regularity in periods of uptrends and downtrends are selected as perceptually important events to improve time series forecasting models. An experimental evaluation based on price prediction on ten corn futures contracts (derivatives) provides evidence that the proposed approach is promising.**

***Index Terms*—time series forecasting, agribusiness, risk management, news events, machine learning.**

I. Introdução

A cadeia produtiva do agronegócio é considerada complexa devido aos muitos fatores envolvidos [1], como efeitos climáticos, controle de pragas e doenças, controle de insumos (*e.g.* fertilizantes, sementes, defensivos, vacinas e máquinas), serviços de crédito (*e.g.* câmbio e taxas de juros), nas atividades de pré-produção e produção de agricultura e pecuária propriamente ditas; e nas etapas pós-produção, como embalagem, armazenagem, industrialização, transporte e distribuição. Esta diversidade de fatores é responsável pelo conceito de riscos e incertezas da cadeia produtiva do agronegócio, com a principal consequência de que o preço de venda durante a etapa de comercialização foge ao controle do produtor. Dependendo da oscilação desses preços, o produtor pode não cobrir seus custos ou não obter a margem de lucro esperada, dificultando honrar compromissos adquiridos. Assim, uma gestão moderna do agronegócio incorpora instrumentos para gerenciamento de riscos com o objetivo de amenizar tais incertezas e conferir um estabilidade mínima ao produtor [2].

O mercado de futuros é um dos principais instrumentos para o gerenciamento de riscos no agronegócio [3]. De forma geral, produtores podem comprar insumos ou vender sua produção visando fixar o preço desejado antecipadamente e se proteger contra oscilações. Tal instrumento, também chamado de derivativos, é gerenciado por meio de bolsas de mercadorias, automatizado por sistemas computacionais, com regras claras e transparentes aos envolvidos. Nesse sentido, uma das tarefas cruciais no gerenciamento de riscos é a previsão de preços de produtos de agronegócios, fator importante para todo o processo de tomada de decisão [2], [1].

Existem vários métodos computacionais para apoiar a previsão de preços [4]. Em geral, podem ser organizados em métodos estatísticos, como o ARIMA e Holt-Winters, e métodos baseados em aprendizado de máquina, como redes neurais, máquinas de vetores de suporte e técnicas dos vizinhos mais próximos. Um fator em comum a todos esses métodos é a construção de um modelo de previsão realizada a partir de uma série temporal histórica, ou seja, a partir da análise de comportamento passado e da extração de padrões dessa série temporal [4].

Estudos recentes demonstram que a previsão de séries temporais pode ser melhorada de forma significativa ao considerar informação adicional sobre o problema [5], [6], [7]. Em especial, além da série temporal histórica, também pode ser utilizado o conhecimento textual extraído de notícias, redes sociais, boletins e outras fontes de dados públicas disponíveis

na *web*. Por exemplo, notícias sobre ocorrência de pragas e doenças, queimadas, novas variedades, supersafras ou quebra de safras, linhas de crédito e incentivos fiscais representam um conhecimento valioso sobre o domínio do agronegócio que pode reduzir o erro da previsão.

Um importante desafio de pesquisa está relacionado em como incorporar de forma efetiva este conhecimento textual adicional durante a construção do modelo de previsão [6], [8]. Os trabalhos existentes consideram que tal informação adicional seja previamente rotulada por um especialista do domínio, como em notícias positivas ou negativas, ou consideram a existência de um dicionário de palavras-chave relacionadas com quedas e altas dos preços. No entanto, assumir a existência desta informação rotulada ou dicionários sobre o domínio é praticamente inviável em aplicações do mundo real. Primeiro, é difícil até mesmo para os especialistas de domínio rotularem uma amostra significativa de informação para identificar o impacto (positivo ou negativo) na cotação do preço, uma vez que a oscilação de preços é ocasionada por um grande conjunto de fatores. Dado que o conhecimento é dinâmico e com alta taxa de atualização, a rotulação frequente dessas informações é um processo muitas vezes impraticável.

Neste trabalho é apresentada uma abordagem de previsão de séries temporais que permite incorporar conhecimento externo na forma de eventos extraídos de notícias de agronegócios, sem a necessidade de previamente rotular a informação textual. Eventos são informações factuais extraídas de notícias e boletins informativos, sendo comumente definidos como "algo específico que ocorre em determinado tempo e lugar" [9]. Por exemplo, o texto "10/04/2017 - Excesso de chuva interfere na qualidade das lavouras de milho em Mato Grosso" é um evento que possui as componentes de tempo, localidade e o fato relacionado, como causa e efeito. Embora existam vários trabalhos relacionados à extração de eventos em notícias, geralmente envolvendo técnicas de processamento de linguagem natural, a incorporação automática de eventos para melhorar a tarefa de previsão em séries temporais é pouco explorada — mesmo sendo um tipo de informação potencialmente útil para gerenciamento de riscos em agronegócios. Nesse sentido, as principais contribuições desse trabalho são listadas a seguir:

- É apresentada uma proposta para identificação de eventos perceptualmente importantes para a tarefa preditiva. Nesse caso, são identificados automaticamente períodos de altas e baixas significativas de uma série temporal, conhecidos na literatura como pontos perceptualmente importantes (PIP). Os eventos publicados nestes períodos são considerados candidatos para apoiar a previsão. No entanto, são selecionados apenas os eventos denominados PIP-recorrentes, ou seja, que possuem no mínimo dois eventos vizinhos publicados em períodos de PIP dentre os seus *top-t* eventos mais próximos. Assim, eventos similares publicados com certa regularidade em períodos de altas e baixas são automaticamente selecionados para serem incorporados na previsão de séries temporais.
- É apresentada uma extensão de um algoritmo de aprendizado de máquina para previsão em séries temporais baseado na técnica k-subséries mais próximas proposto em [7], com o diferencial de utilizar os eventos perceptualmente importantes para ajustar a previsão. Nesse caso, o valor predito é baseado nas k subséries mais similares da subsérie atual, ou seja, busca-se padrões de observações históricas com o mesmo comportamento das observações mais recentes. Na abordagem aqui proposta, além da tradicional correlação entre as subséries, o critério de proximidade entre subséries considera também a proximidade semântica entre eventos perceptualmente importantes publicados nos períodos das subséries.

Foi realizada uma avaliação experimental envolvendo gerenciamento de riscos de preços no agronegócio, em particular, na previsão do valor de encerramento de contratos futuros de Milho negociados na BM&F-BOVESPA. Foi utilizada uma série temporal composta por 10 anos de cotações do milho, bem como milhares de eventos sobre o domínio coletados em diversas agências e portais especializados. O experimento envolveu a previsão do valor de fechamento de dez contratos de milho entre 2016 e 2017. A proposta foi comparada com o método kNN-TSP, considerado um dos métodos estado-da-arte na previsão de séries temporais [10]. Os resultados experimentais revelam que a abordagem proposta reduziu o erro de previsão de três contratos futuros (dentre dez contratos) em relação ao método kNN-TSP, além de manter o mesmo poder preditivo nos outros sete contratos. Os resultados experimentais indicam que a incorporação de eventos é potencialmente útil para o gerenciamento de riscos no agronegócio, bem como também permite analisar e interpretar em um maior nível semântico os fatores envolvidos nas previsões.

O restante deste trabalho está organizado da seguinte maneira. Na próxima seção são apresentados conceitos e fundamentos básicos, bem como trabalhos relacionados na literatura. Na Seção III são discutidos os detalhes da proposta deste trabalho. A avaliação experimental e análise dos resultados são apresentadas na Seção IV. Por fim, limitações desse estudo e direções para trabalhos futuros são apresentadas na Seção V.

II. Fundamentos e Trabalhos Relacionados

A negociação de contratos futuros de produtos agrícolas, também chamados de derivativos de *commodities*, representam uma das principais estratégias para gerenciamento de riscos no agronegócio [2]. Um contrato futuro indica o compromisso de comprar ou vender determinado produto numa data futura, por um preço previamente fixado entre a partes. O principal objetivo deste tipo de contrato é a proteção dos envolvidos, como agricultores, pecuaristas, comerciantes, industriais, instituições financeiras e investidores, contra as oscilações dos preços desses produtos agrícolas. Assim, produtores podem negociar com base em análises do mercado e minimizar os riscos de prejuízos.

Para exemplificar, considere um produtor de frangos que utiliza produtos derivados de milho como insumo para alimentação das aves. Nesse caso, altas inesperadas nos preços do milho podem afetar diretamente na produção e no lucro estimado. Para se proteger desta oscilação, o produtor faz um

pedido de compra antecipada de milho, por meio de contrato futuro ao preço de R$ 15,00 por saca para entrega em 120 dias, por exemplo. Se neste período ocorrer um aumento inesperado para R$19,00, o produtor receberá R$ 4,00 de bônus por saca. Do mesmo modo, se ocorrer uma queda do preço para R$ 12,00 por saca, o produtor deve pagar ao mercado R$ 3,00 por saca. Em todos os cenários, o produtor na prática negocia seu produto no valor de R$ 15,00, conforme estipulado no início do contrato. O mecanismo de compensação é realizado de forma automática, geralmente por ajustes diários, utilizando margens de garantias que ambos compradores e vendedores devem depositar junto às suas corretoras. Assim, é possível afirmar que este tipo de gerenciamento de riscos funciona como uma espécie de seguro aos envolvidos, no qual agentes especuladores de mercado assumem o risco da oscilação dos preços.

Uma das principais tarefas no gerenciamento de riscos de preços no agronegócio é a previsão do valor do contrato futuro para otimizar o lucro dos envolvidos. A análise de séries temporais com a cotação histórica dos preços é uma abordagem muito utilizada pela indústria e que tem obtido cada vez mais atenção nos últimos anos, conforme o avanço de métodos de aprendizado de máquina [11]. A ideia central é identificar tendências e sazonalidades na série temporal que podem ser utilizadas na previsão de novas observações, conforme discutido nas próximas seções.

A. Previsão em Séries Temporais

As séries temporais podem ser descritas como uma sequência ordenada de observações [12]. Uma série temporal de tamanho n é definida como $X = (x_1, x_2, ..., x_n)$ na qual $x_t \in \Re$ representa uma observação no instante de tempo t. A série temporal pode ser determinística ou estocástica [13]. Uma série é dita determinística quando seus dados são representados por uma função em termo do tempo $y = f(tempo)$, ou seja, possui um comportamento regular e previsível. Já uma série estocástica possui um termo adicional ϵ, representado pela função $y = f(tempo, \epsilon)$, que é responsável por produzir uma série de comportamento não regular. Muitas aplicações práticas são baseadas em séries temporais estocásticas, sendo necessários métodos computacionais mais robustos para realização da tarefa preditiva [11].

Os métodos de previsão em séries temporais podem ser estatísticos ou baseados em aprendizado de máquina. Por décadas, os métodos estatísticos foram considerados o estado-da-arte nesse tipo de tarefas, em particular, métodos baseados em médias móveis e autoregressão. Em geral, o objetivo é identificar os coeficientes de um modelo de forma a ajustar uma função aos dados da série [14]. Um dos métodos estatísticos mais utilizados na literatura é o ARIMA (Modelos Autorregressivos Integrados de Médias Móveis) [4], que considera tanto a autocorrelação entre observações do passado e futuro, bem como médias móveis para identificação da tendência. A principal característica dos métodos estatísticos é que são paramétricos, ou seja, é necessário assumir *a priori* que as observações seguem uma determinada distribuição. Por outro lado, esta característica é também uma limitação desses modelos, pois exige conhecimento de um especialista, tanto sobre domínio da aplicação quanto sobre métodos computacionais.

Dentre os métodos baseados em aprendizado de máquina para previsão em séries temporais, destaca-se o uso de Redes Neurais (*MLP - Multilayer Perceptron*), Máquinas de Vetores de Suporte (SVM) e dos k-Vizinhos mais Próximos (kNN-TSP) [15], [12]. Um ponto em comum a esses métodos é que a série temporal é dividida em um conjunto de subséries. As subséries são utilizadas como conjunto de treinamento para o aprendizado, geralmente, utilizando uma estratégia de regressão, em que o atributo classe da tarefa de aprendizado é um valor numérico. Por exemplo, dada uma subsequência da série, seu atributo classe é o próximo valor dessa subsequência.

Neste trabalho há maior interesse em métodos de previsão baseados no k-Vizinhos mais Próximos (kNN-TSP). É uma estratégia simples e intuitiva, que obtém resultados promissores em muitas aplicações [16]. A premissa básica é que as k subsequências passadas que possuem comportamento mais similar à subsequência mais recente da série temporal podem ser utilizadas para prever o valor da próxima observação. Assim, seja uma série temporal $X = (x_1, x_2, ..., x_n)$, a observação x_{n+h} indica o valor a ser predito, em que h é o horizonte de previsão. Nesse caso, se $h = 1$, então será predito o valor seguinte da série.

De forma geral, a previsão utilizando kNN-TSP utiliza a subsérie atual $Q = (x_{n-r+1}, x_{n-r+2}, ..., x_n)$ de tamanho r, em que $r \in \mathbb{N}$ é definido conforme o domínio da aplicação ou o estimado durante o próprio processo de análise experimental. Seja $S = (S^{(1)}, ..., S^{(k)})$ as k subséries mais próximas à subsérie Q, em que $S^{(j)} = (x_{ini}, x_{ini+1}, ..., x_{end})$ com $ini \geq 1$, $end \leq (n - 2r + 1)$ e $end - ini = r$. Seja $S^{(j)}_{+h}$ uma notação que indica o valor da próxima observação conhecida da subsérie $S^{(j)}$ considerando um horizonte de previsão h. A previsão é realizada por meio de uma função de previsão $f(S)$, descrita na Equação (1), que retorna o valor x_{n+h} considerando o erro médio entre as últimas observações de cada subsérie do passado e da subsérie atual (primeiro termo da soma), ajustada com a última observação da subsérie atual (segundo termo da soma) [17]. Tal função de previsão permite "extrapolar" os valores preditos fora do intervalo do conjunto de treino, o que possibilita sua aplicação para séries temporais com diversos comportamentos.

$$x_{n+h} = f(S, h) = (\frac{1}{k} \sum_{j=1}^{k} (S^{(j)}_{+h} - S^{(q)}_{r})) + S^{(q)}_{r} \qquad (1)$$

A aplicação deste método de previsão baseada em subséries mais próximas é ilustrado pela Figura 1. Neste exemplo, três subséries mais similares à subsérie mais recente foram identificadas. Os pontos circulares indicam a próxima observação $S^{(j)}_{+h}$ de cada subsérie vizinha. O ponto no quadrado destacado indica o valor a ser predito.

O ponto crucial do método de previsão baseado no kNN-TSP é o critério usado para identificar a proximidade entre

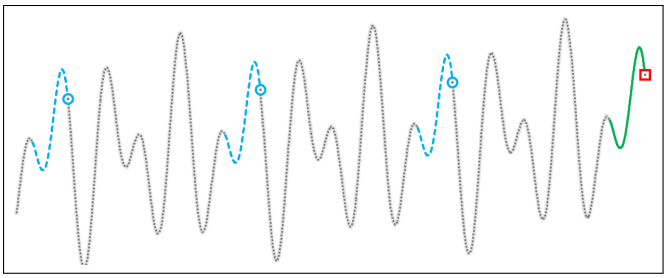

Figura 1. Ilustração exemplificando o processo de previsão com o método k-NN. Em azul são identificadas as três subséries mais similares à subsérie mais recente (em vermelho). Os pontos circulares indicam a próxima observação $S^{(j)}_{+h}$ de cada subsérie vizinha. O ponto no quadrado destacado indica o valor a ser predito (Fonte: [18]).

duas subséries, discutido a seguir.

B. O problema do cálculo da proximidade entre subséries

O critério mais comum da literatura para identificar a proximidade entre duas subséries é a distância Euclidiana, definida na Equação 2 para duas subséries $Q = (q_1, ..., q_r)$ e $C = (c_1, ..., c_r)$, com r observações cada.

$$euc(Q, C) = \sqrt{\sum_{i=1}^{r}(q_i - c_i)^2} \tag{2}$$

A distância euclidiana consiste no alinhamento linear entre cada subsérie. No entanto, em muitos cenários práticos é interessante utilizar medidas de proximidade entre subséries que permitem um alinhamento não linear. Nesse sentido, a distância DTW (*Dymamic Time Warping*) [19] é umas das mais efetivas quando aplicadas para buscar similaridades entre subsequencias de uma série temporal. O DTW permite distorções da série em relação à linha temporal, reajustando o alinhamento dos pontos de forma mais significativa [10]. Na Figura 2 é ilustrada uma comparação entre ambas as medidas para explicitar a diferença que entre alinhamento linear e não linear.

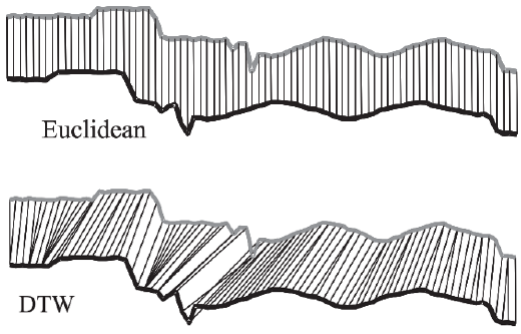

Figura 2. Comparação entre as distâncias Euclidiana (alinhamento linear) e DTW (alinhamento não linear) para cálculo de proximidade entre subséries (Fonte: [20]).

O cálculo da distância DTW é dividido em duas partes. A primeira etapa é responsável por calcular uma matriz de custo $A_{r\times r}$ entre as subséries Q e C. Nessa matriz, cada posição A_{ij} é obtida pela distância euclidana entre os valores das observações q_i e c_j, ou seja, $A_{ij} = (q_i - c_j)^2$.

O alinhamento não linear realizado pelo DTW consiste na obtenção da distância de menor custo entre a posição de origem A_{11} até A_{rr}. Na Equação 3 é apresentada a relação de recorrência do DTW para obtenção do caminho de menor custo.

$$\begin{aligned} DTW(Q,C) = d(q_i, c_j) + min\{ & \\ & DTW(q_i - 1, c_j - 1), \\ & DTW(q_i - 1, c_j), \\ & DTW(q_i, c_j - 1) \\ \} & \end{aligned} \tag{3}$$

Na Figura 3 é ilustrado um exemplo de caminho de menor custo. Observe que quando o caminho de menor custo é representado estritamente pela diagonal da matriz $A_{r\times r}$, então a DTW é equivalente à distância euclidiana; uma vez que foi realizado um alinhamento linear. Qualquer caminho de menor custo diferente da diagonal representa uma solução em que uma distorção no tempo foi necessária para obter uma melhor proximidade entre as duas subséries.

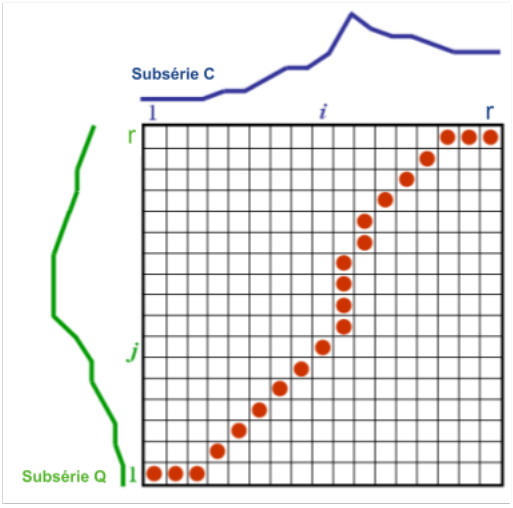

Figura 3. Matriz de custo DTW e caminho de menor custo (Adaptado de [21]).

Embora a distância DTW e Euclidiana sejam reportadas como as mais utilizadas na previsão de séries temporais com o método k-NN, ambas possuem a limitação de serem baseadas apenas nas observações históricas das séries temporais. Conforme discutido na introdução desse trabalho, há diversos fatores externos que podem afetar a proximidade das subséries, em especial, eventos extraídos de notícias publicadas nos períodos de tempo dessas subséries. Dessa forma, uma proposta que permita estender o cálculo da proximidade para considerar tal informação externa é potencialmente promissora para tarefas de análise preditiva em séries temporais.

III. Abordagem Proposta: Eventos Perceptualmente Importantes na previsão em Séries Temporais

Neste trabalho é apresentada uma abordagem para incorporar informações externas às observações da série temporal com a finalidade de melhorar a previsão de valores futuros, em particular, aplicado ao gerenciamento de riscos em agronegócios. São utilizados eventos extraídos de notícias e boletins publicados em diversos portais e agências brasileiras sobre o tema como informação externa.

Para minimizar o esforço de um especialista do domínio informar quais os eventos são interessantes para a tarefa preditiva, a abordagem aqui proposta identifica eventos perceptualmente importantes com base em períodos críticos da série temporal (pontos significativos de alta e baixa da cotação do produto agrícola).

Os eventos são informações factuais relevantes para o domínio, pois geralmente indicam lugares de ocorrência, organizações envolvidas, causas e efeitos. Na prática, é analisado se os eventos passados extraídos das notícias possuem semelhanças com os eventos atuais para ajustar o alinhamento não linear realizado pelo distância DTW. Assim, é proposta uma modificação no cálculo da matriz de custo do DTW e analisado o efeito da incorporação dos eventos na previsão de séries temporais.

Uma visão geral da abordagem proposta é apresentada na Figura 4. Nas próxima seções, cada passo da abordagem é descrito em mais detalhes.

A. Definição de Série Temporal, Coleta e Pré-processamento de Eventos

A abordagem recebe como entrada uma série temporal do domínio de agronegócio e um conjunto de palavras-chave sobre o domínio. Tais palavras-chave são utilizadas para a coleta de uma amostra de eventos relacionados a partir de uma base de conhecimento. Por exemplo, neste trabalho utilizou-se a base de conhecimento do projeto *Websensors*[1], uma ferramenta online compostas por diversos *crawlers* para sites especializados visando garantir a qualidade dos dados, que são disponibilizados para a finalidade de extração de conhecimento [22]. A definição da palavra-chave não exige um conhecimento aprofundado sobre o tema, uma vez que é utilizada apenas para construção do conjunto de eventos relacionados. Por exemplo, neste trabalho foi utilizada a palavra-chave "milho" para a coleta de eventos.

Como parte do processo de aprendizado e seleção de um modelo para previsão de séries temporais, a série de entrada é dividida em duas séries temporais: treinamento e teste (Passo 1). A série temporal de treinamento é utilizada como entrada para o método kNN-TSP. A série de teste é utilizada para avaliar a eficácia do método considerando suas configurações. Na prática, há um processo de refinamento para buscar a melhor configuração do kNN-TSP em relação à previsão na série temporal de teste. Uma vez obtida tal configuração, então o modelo final é utilizado para realizar a tarefa preditiva em dados não vistos.

No Passo 2 da abordagem é realizado um pré-processamento dos eventos coletados a partir da(s) palavra(s)-chave fornecida(s). Geralmente os dados textuais possuem várias formatações, sendo necessária a padronização dos textos, na qual são convertidos em texto plano. Um problema inicial deste processo são os milhares de termos redundantes e desnecessários, o que tornaria o processo de extração de conhecimento dos textos de baixa qualidade. Assim, é realizado um processo de seleção dos termos mais representativos possíveis, que envolve (i) a eliminação de stopwords, que são termos não representativos para a extração e que são artigos, pronomes e advérbios e a (ii) *stemmização* dos termos que reduz cada palavra à sua forma básica sem inflexões, ou seja, removem-se todos os sufixos que a variam morfologicamente.

O final do processo de seleção de termos é a construção de uma representação no modelo espaço-vetorial denominada *bag-of-words*. Também é conhecida como tabela atributo-valor (Figura 5) que indica os termos e sua importância por evento. Como eventos são compostos por textos curtos, o valor das células indica a presença ou ausência de um termo no evento.

B. Identificação de Pontos Perceptualmente Importantes

No Passo 3 da abordagem é realizada a identificação de pontos perceptualmente importantes na série temporal, denominados PIP's (*Perceptually Important Points*) [23]. Um PIP pode ser definido como um ponto crítico de uma série temporal, uma vez que representa pontos com valores muitos diferentes de seus pontos vizinhos, geralmente representados por picos ou vales de uma série.

A identificação dos PIPs na série temporal é realizada de forma automática por meio de um algoritmo baseado na distância euclidiana entre pontos [24]. Na inicialização do algoritmo de identificação de PIPs, o primeiro e último ponto da série são definidos como PIPs iniciais. Em seguida, para cada dois PIPs identificados, calcula-se o ponto mais distante entre eles, considerando a distância euclidiana. Esse ponto mais distante é definido como um novo PIP. Este processo de extração é executado recursivamente até que se tenha a quantidade de PIPs desejados. Um exemplo da execução do algoritmo de identificação PIPs é ilustrado na Figura 6.

C. Identificação de Eventos Perceptualmente Importantes

Ainda como parte do Passo 3, é realizado um processo para identificar os eventos mais relevantes para a previsão da série temporal. Nesta abordagem, adotou-se uma estratégia que consiste em filtrar os eventos publicados nos períodos dos pontos perceptualmente importantes. Dessa forma, uma nova tabela atributo-valor é obtida utilizando apenas os eventos publicados nos pontos críticos da série temporal.

Os eventos perceptualmente importantes são obtidos por meio do conceito de evento PIP-recorrente, proposto neste trabalho. Um evento é PIP-recorrente se o seu conteúdo ocorre com certa frequência em mais de um período de PIP. Assim, considerando o conjunto de todos os eventos obtidos

[1]Projeto Websensors: http://www.websensors.net.br/

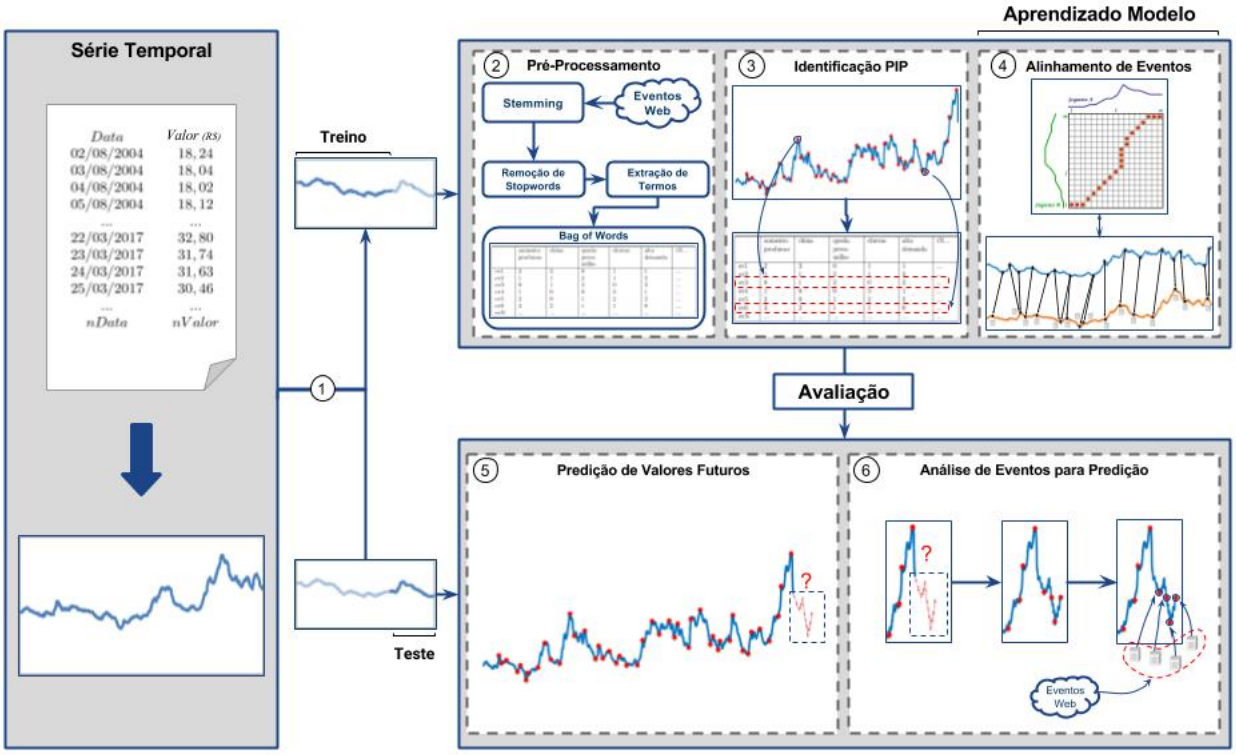

Figura 4. Visão geral da abordagem proposta para incorporação de eventos durante a previsão de séries temporais.

	t_1	t_2	...	t_M
e_1	a_{11}	a_{12}	...	a_{1M}
e_2	a_{21}	a_{22}	...	a_{2M}
⋮	⋮	⋮	⋱	⋮
e_N	a_{N1}	a_{N2}	...	a_{NM}

Figura 5. Representação dos eventos por meio de uma tabela atributo-valor.

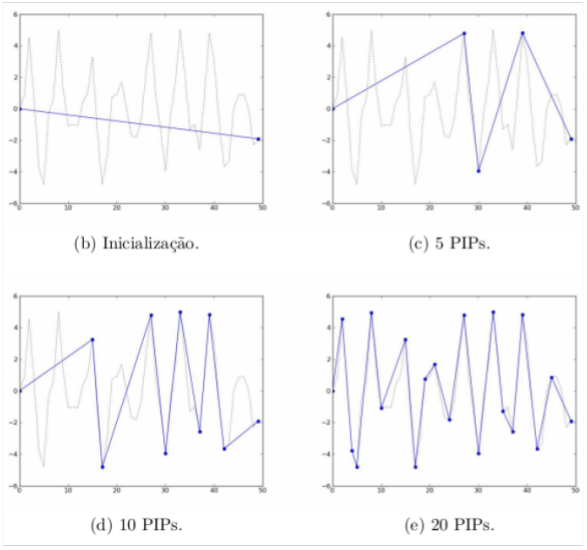

Figura 6. Exemplo da execução do algoritmo de identificação de PIP's em uma série temporal (Fonte: [25]).

no Passo 2 e os eventos filtrados pelos períodos de PIPs, os eventos perceptualmente importantes são aqueles eventos que possuem pelo menos dois eventos vizinhos publicados em períodos de PIP dentre os seus $top-t$ eventos mais próximos. Essa estratégia de identificação de eventos perceptualmente importantes tem o propósito de eliminar eventos publicados ocasionalmente, sem relação semântica com períodos críticos (altas e baixas) da série temporal. Ainda, espera-se que eventos com conteúdo frequentemente publicado nos pontos de alta e baixa possam ser úteis para a previsão, sem a necessidade de uma seleção manual de eventos por um especialista do domínio.

A proximidade entre dois eventos depende de medidas apropriadas para similaridade entre dados textuais. Neste trabalho foi utilizada a medida de similaridade de Jaccard, útil para dados binários. Na Equação 4 é apresentada a medida de dissimilaridade baseada no Jaccard, que calcula a quantidade de informação (termos) compartilhada entre dois eventos E_i e E_j.

$$djac(E_i, E_j) = 1 - \frac{E_i \cap E_j}{E_i \cup E_j} \quad (4)$$

D. Alinhamento de Eventos

Com a extração dos eventos perceptualmente importantes, é possível enriquecer o critério de distância DTW. No Passo 4, o alinhamento não linear de eventos é ajustado considerando um peso para a informação textual dos eventos.

A extensão realizada no DTW é baseada no cálculo da matriz de custo. Seja $A^{ts}_{r\times r}$ a matriz de custo entre duas subséries Q e C (via distância euclidiana). Seja $A^{ev}_{r\times r}$ a matriz de custo entre os eventos publicados nos períodos de tempo das subséries Q e C (via dissimilaridade de Jaccard). A matriz de custo final é obtida conforme a Equação 5, em que α indica o peso dos eventos na matriz de custo final ($0 \leq \alpha \leq 1$). Ainda, a notação $\bar{A}_{ij}$ representa uma versão normalizada dos valores da matriz, permitindo que a combinação linear entre as duas matrizes respeite a mesma escala numérica.

$$A_{r\times r} = \sum_{i=1}^{r}\sum_{j=1}^{r}(1-\alpha)\bar{A}^{ts}_{ij} + \alpha\bar{A}^{ev}_{ij} \quad (5)$$

Uma vez definida a matriz de custo final, o cálculo do caminho de menor custo do DTW é obtido da forma original. A abordagem proposta pode ser vista como uma generalização do DTW tradicional, uma vez que $\alpha = 0$ anula a importância do conhecimento externo. Na abordagem proposta, o valor de α é obtido automaticamente dentro da própria procedimento de treinamento e teste utilizado para seleção do melhor modelo de previsão.

E. Previsão na Série Temporal

O procedimento de previsão (Passos 5 e 6) é baseado na mesma estratégia do método kNN-TSP, descrito anteriormente. O horizonte de previsão h é definido pelo usuário ou por algum requisito da aplicação. Para valores com $h > 1$ é utilizada um critério de previsão por atualização, em que um valor predito é adicionado no final da série temporal e este valor é considerado parte da série temporal para previsão do próximo valor (e assim sucessivamente). Como ocorre em qualquer método, a previsão de horizontes muito grandes é um problema complexo devido à propagação de erro das predições anteriores.

Na abordagem proposta, o conhecimento adicional sobre o domínio representado pelos eventos perceptualmente importantes pode ser utilizado como uma estratégia descritiva do modelo. Assim, ao realizar uma previsão é possível também apresentar ao usuário quais os eventos perceptualmente importantes (atuais e do passado) que foram empregados para ajustar a matriz de custo do DTW. Na prática, a abordagem proposta tem a vantagem de permitir a interpretação/exploração da tarefa preditiva — ao contrário da maioria dos métodos existentes que são considerados um modelo de previsão caixa preta.

IV. Avaliação Experimental

A avaliação experimental realizada neste trabalho envolve a área de gerenciamento de riscos de preço para contratos de Milho negociados na BM&F-Bovespa. Tal produto agrícola possui contratos futuros com vencimentos em sete meses de cada ano (janeiro, março, maio, julho, agosto, setembro e novembro), com data de vencimento no dia 15 de cada mês (ou no próximo dia útil)[2]. O objetivo geral da avaliação experimental é analisar o impacto da incorporação de eventos perceptualmente importantes (PIP-recorrentes) na previsão de valores de contratos futuros em comparação com o método kNN-TSP, que só utiliza a série temporal histórica.

Foram utilizados dez contratos futuros (Jan-2016, Mar-2016, Mai-2016, Jul-2016, Ago-2016, Set-2016, Nov-2016, Jan-2017, Mar-2017 e Mai-2017) para uma simulação envolvendo previsão de preços. Na Figura 7 é apresentada a série histórica da cotação do preço do milho, com granularidade mensal, e os pontos realçados em vermelho indicam os meses de vencimento de cada contrato.

De acordo com a literatura envolvendo gerenciamento de riscos de preços, parte da produção de milho é negociada entre três a seis meses antes da colheita como mecanismo de proteção à variações inesperadas de preço [26]. Nesta avaliação experimental, o horizonte de previsão foi de seis meses ($h = 6$) para cada contrato futuro. O número máximo de PIPs foi de 100 em todos os cenários analisados e o número de *top-t* eventos vizinhos foi de 30. Nessa configuração, a base de conhecimento extraída do projeto Websensors foi composta de 25.857 eventos identificados como perceptualmente importantes (PIP-recorrentes) com um total de 6.603 termos após o processo de pré-processamento de textos. Na Figura 8 é apresentada um exemplo da interface desenvolvida para seleção de eventos relevantes ao domínio. Na Figura 9 é apresentada a quantidade de eventos perceptualmente importantes para cada mês.

Para avaliar o efeito do uso de eventos na previsão, o método base kNN-TSP foi avaliado com diferentes configurações. O número de vizinho k foi testado no intervalo $[1, 30]$. Foi utilizada a distância DTW para identificar as k subséries mais próximas. O tamanho mínimo da subsérie foi definido como $r = 6$ e o máximo é de $r = 18$. No caso da abordagem proposta, além dos parâmetros anteriores foi analisado o parâmetro α no intervalo $[0, 1]$, que identifica o peso dos eventos durante a previsão.

O critério de avaliação utilizado para estimar o erro da previsão foi o MAPE (*Mean absolute percentage error*) [27], conforme definido na Equação 6. Nesta equação, h é o horizonte de previsão, $real_t$ é o valor real observado e $pred_t$ é o valor predito pelo método. Em termos práticos, o MAPE é uma medida de percentagem de erro que, em uma simulação, indica o quão próximo a previsão foi realizada em relação aos valores conhecidos da série temporal.

$$MAPE = \frac{100}{h}\sum_{t=1}^{h}\left|\frac{real_t - pred_t}{real_t}\right| \quad (6)$$

Diversos modelos de previsão foram construídos conside-

[2]Informações sobre contrato futuro de milho: http://www.bmfbovespa.com.br/pt_br/produtos/listados-a-vista-e-derivativos/commodities/futuro-de-base-de-preco-de-milho.htm

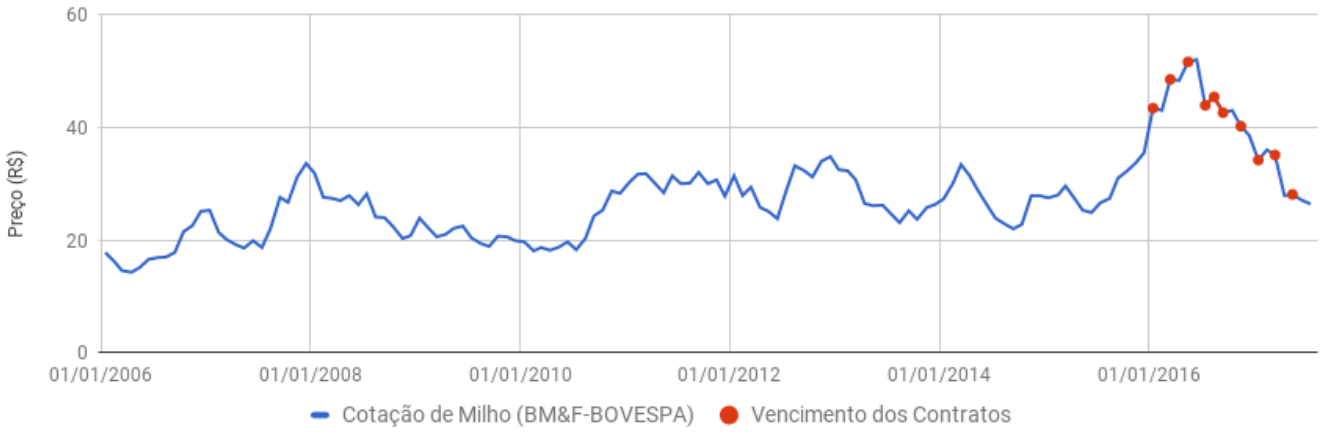

Figura 7. Série de cotação de preços de milho de Jan-2016 até Jul-2017. As observações destacadas em vermelho representam meses de vencimento de contratos utilizados na avaliação experimental.

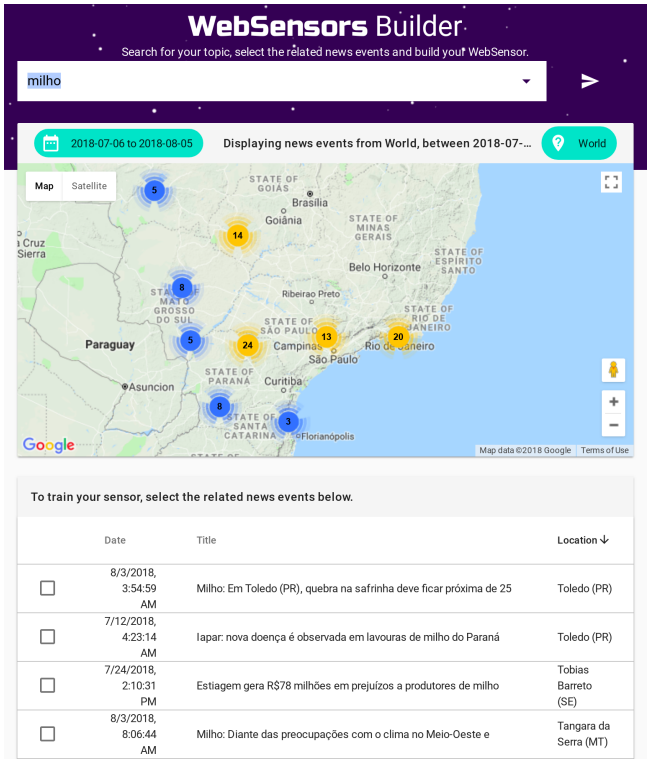

Figura 8. Interface para selecionar amostra de eventos para um determinado domínio.

rando a série temporal de treinamento por meio da variação dos parâmetros. A configuração que minimiza o valor MAPE em cada contrato foi selecionada para analisar a eficácia da previsão no cenário simulado.

Na Tabela I é apresentada uma visão geral dos resultados experimentais. É possível notar que a incorporação de eventos perceptualmente importantes permitiu a redução do MAPE em três contratos (Ago-216, Set-2016 e Nov-2016). Nos outros contratos, a melhor previsão foi obtida com o parâmetro $\alpha = 0$, que anula o efeito dos eventos perceptualmente importantes.

Tabela I
COMPARAÇÃO GERAL DO ERRO DE PREVISÃO (MAPE) ENTRE A ABORDAGEM PROPOSTA E O MÉTODO KNN-TSP

Contrato	KNN-TSP			PROPOSTA			
	k	w	MAPE (%)	k	w	α	MAPE (%)
Jan-2016	2	9	7.38	2	9	0	7.38
Mar-2016	1	9	15.13	1	9	0	15.13
Mai-2016	2	12	25.16	2	12	0	25.16
Jul-2016	1	16	12.38	1	16	0	12.38
Ago-2016	10	18	16.13	2	6	0.95	14.10
Set-2016	8	18	4.76	1	16	0.85	4.49
Nov-2016	2	13	2.98	3	14	0.75	2.70
Jan-2017	10	18	3.61	10	18	0	3.61
Mar-2017	1	9	15.13	1	9	0	15.13
Mai-2017	2	12	25.16	2	12	0	25.16

A abordagem proposta tem um efeito de ajuste da previsão de um método base (no caso o kNN-TSP). Assim, no pior caso, o profissional da área de gerenciamento de riscos utilizará as predições dos métodos bases. Porém, em cenários em que há eventos relevantes para a previsão, então o ajuste pode ser benéfico. Em especial, em ambientes envolvendo análise/estudos das variáveis que estão afetando o preço do produto agrícola, é possível inclusive identificar quais eventos que foram utilizados para realizar o ajuste para fornecer uma interpretação da previsão.

V. CONSIDERAÇÕES FINAIS

Neste trabalho foi proposta uma abordagem para incorporação de informação externa na previsão de séries temporais. A informação externa é representada por meio de eventos, em especial, eventos perceptualmente importantes que foram publicados em períodos de alta e baixa da série temporal. A aplicação de interesse deste trabalho é o gerenciamento de riscos em agronegócios, ou seja, identificar eventos perceptualmente importantes que podem afetar o preço futuro da cotação de um produto agrícola.

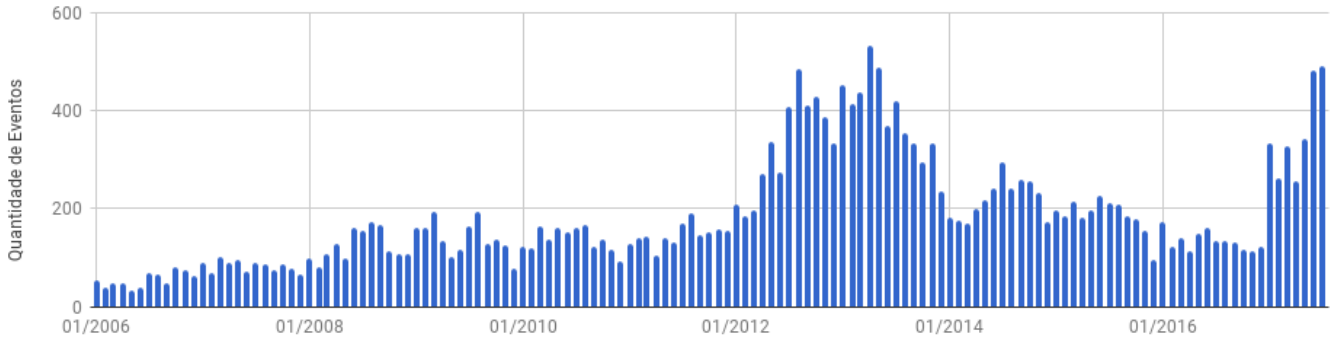

Figura 9. Quantidade de eventos perceptualmente importantes identificados em cada mês considerado 100 PIPs.

Foi realizada uma análise experimental envolvendo contratos futuros de Milho negociados na BM&F-BOVESPA. A abordagem proposta permitiu reduzir o erro da previsão em três de dez contratos futuros. Embora seja um indicativo de que eventos perceptualmente importantes possam ser úteis para esta tarefa, ainda há algumas limitações da análise experimental realizada e que são direções para trabalhos futuros:

- A quantidade de eventos perceptualmente importantes é relacionada ao número de PIPs da série temporal. Esse número foi mantido constante durante a análise experimental (100 PIPs e 30 eventos vizinhos). No entanto, tais parâmetros também podem ser variados automaticamente dentro da própia estratégia de seleção dos parâmetros do modelo; e
- A abordagem proposta é interessante na perspectiva de um sistema de *Data Analytics* para apoio à tomada de decisão. Assim, os usuários poderam analisar potenciais eventos que afetam a cotação do preço. No entanto, ainda é necessário realizar o desenvolvimento de ferramentas de visualização de dados apropriadas para exploração desses resultados.

Outras direções para trabalhos futuros envolvem o estudo de métodos para incorporar informações externa de eventos perceptualmente importantes em outros algoritmos de previsão de séries temporais, como Redes Neurais e Máquinas de Vetores de Suporte. Além disso, iniciativas envolvendo aprendizado semissupervisionado também estão sendo planejadas, uma vez que usuários podem oferecer *feedback* em um pequeno conjunto de eventos de interesse para melhorar a tarefa de previsão.

Agradecimentos

Este trabalho contou com o apoio das seguintes agências de fomento: FAPESP (Processo 2017/08804-2), Fundect-MS (Processo 14/08996-0), CAPES, CNPq e FINEP. Os autores agradecem a NVIDIA pela doação de GPUs (*GPU Grant Program*). Os autores também agradecem ao doutorando Antonio Rafael Sabino Parmezan (ICMC/USP) pelo apoio no desenvolvimento deste trabalho.

Referências

[1] L. Zuin and T. Queiroz, *Agronegócios - Gestão Inovação e Sustentabilidade*. Editora Saraiva, 2015.
[2] F. Schouchana, *Gestão de riscos no agronegócio*. Editora FGV, 2015.
[3] J. Rocha, W. Freire, M. V. L. Bittencourt, and M. C. P. Ribeiro, "Análise das características dos contratos no agronegócio do brasil," *Revista Brasileira de Planejamento e Desenvolvimento*, vol. 4, no. 2, pp. 94–118, 2016.
[4] D. C. Montgomery, L. J. Cheryl, and K. Murat, *Introduction to time series analysis and forecasting*. John Wiley & Sons, 2015.
[5] M. Koppel and I. Shtrimberg, "Good news or bad news? let the market decide," in *Computing attitude and affect in text: Theory and applications*. Springer, 2006, pp. 297–301.
[6] G. Mitra and L. Mitra, *The handbook of news analytics in finance*. John Wiley & Sons, 2011, vol. 596.
[7] R. M. Marcacini, J. C. Carnevali, and J. Domingos, "On combining Websensors and DTW distance for kNN Time Series Forecasting," in *23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 2521–2525.
[8] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," *Expert Systems with Applications*, vol. 41, no. 16, pp. 7653–7670, 2014.
[9] J. Allan, *Topic detection and tracking: event-based information organization*. Springer Science & Business Media, 2012, vol. 12.
[10] A. R. S. Parmezan and G. E. A. P. A. Batista, "A study of the use of complexity measures in the similarity search process adopted by knn algorithm for time series prediction," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, Dec 2015, pp. 45–51.
[11] W. Y. Nyein Naing and Z. Z. Htike, "State of the art machine learning techniques for time series forecasting: A survey," *Advanced Science Letters*, vol. 21, no. 11, pp. 3574–3576, 2015-11-01T00:00:00. [Online]. Available: http://www.ingentaconnect.com/content/asp/asl/2015/00000021/00000011/art00032
[12] C. Chatfield, *The Analysis of Time Series: An Introduction*. Chapman & Hall/CRC Texts in Statistical Science, 2013.
[13] C. Cheng, A. Sa-Ngasoongsong, O. Beyca, T. Le, H. Yang, Z. J. Kong, and S. T. Bukkapatnam, "Time series forecasting for nonlinear and non-stationary processes: a review and comparative study," *IIE Transactions*, vol. 47, no. 10, pp. 1053–1071, 2015. [Online]. Available: http://dx.doi.org/10.1080/0740817X.2014.999180
[14] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.
[15] N. K. Ahmed, A. F. Atiya, N. E. Gayar, and H. El-Shishiny, "An empirical comparison of machine learning models for time series forecasting," *Econometric Reviews*, vol. 29, no. 5-6, pp. 594–621, 2010.
[16] S. Yakowitz, "Nearest-neighbour methods for time series analysis," *Journal of Time Series Analysis*, vol. 8, no. 2, pp. 235–247, 1987. [Online]. Available: http://dx.doi.org/10.1111/j.1467-9892.1987.tb00435.x
[17] C. A. Ferrero, M. C. Monard, H. D. Lee, and F. Wu, "Proposal of a Forecast Function for the KNN algorithm in Time Series (In Portuguese)," *35th Latin American Computing Conference (CLEI)*,

pp. 1–10, 2009. [Online]. Available: http://www.labic.icmc.usp.br/pub/mcmonard/FerreroCLEI09.pdf

[18] A. R. S. Parmezan, "Predição de séries temporais por similaridade," Ph.D. dissertation, Universidade de São Paulo, 2014.

[19] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *KDD Workshop*, 1994.

[20] E. Keogh and A. C. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and Information Systems*, vol. 7, no. 3, pp. 358–386, 2004. [Online]. Available: http://dx.doi.org/10.1007/s10115-004-0154-9

[21] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh, "Searching and mining trillions of time series subsequences under dynamic time warping," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 262–270.

[22] R. M. Marcacini, R. G. Rossi, B. M. Nogueira, L. V. Martins, E. A. Cherman, and S. O. Rezende, "Websensors analytics: Learning to sense the real world using web news events," in *Proceedings of the 23th Brazilian Symposium on Multimedia and the Web - Workshop on Tools and Applications*, 2017, pp. 169–173.

[23] P. E. Tsinaslanidis and D. Kugiumtzis, "A prediction scheme using perceptually important points and dynamic time warping," *Expert Systems with Applications*, vol. 41, no. 15, pp. 6848–6860, 2014.

[24] T.-c. Fu, "A review on time series data mining," *Engineering Applications of Artificial Intelligence*, vol. 24, no. 1, pp. 164–181, 2011.

[25] R. A. Sanches, "Redução de dimensionalidade em séries temporais," Ph.D. dissertation, Universidade de São Paulo, 2006.

[26] E. Batistella, "Comercialização de milho no brasil: análise da utilização do mercado de futuros da bm&f. 2006. 161 f," Ph.D. dissertation, Dissertação (Mestrado Multiinstitucional em Agronegócios)–Universidade Federal de Mato Grosso do Sul.

[27] S. Makridakis, "Accuracy measures: theoretical and practical concerns," *International Journal of Forecasting*, vol. 9, no. 4, pp. 527–529, 1993.